Explanation Closure, Action Closure, and the Sandewall Test Suite for Reasoning about Change

Lenhart K. Schubert

Technical Report 440
October 1992

UNIVERSITY OF
ROCHESTER
COMPUTER SCIENCE

Explanation Closure, Action Closure, and the Sandewall Test Suite for Reasoning about Change*

Lenhart K. Schubert

Department of Computer Science, University of Rochester
Rochester, NY 14627

Abstract. *Explanation closure* (EC) axioms were previously introduced as a means of solving the frame problem. This paper provides a thorough demonstration of the power of EC (combined with *action closure*) for reasoning about dynamic worlds, by way of Sandewall's recently compiled test suite of 12-or-so problems (Sandewall, 1991). Sandewall's problems range from the "Yale turkey shoot" (and variants) to the "stuffy room" problem, and were intended as a test and challenge for nonmonotonic logics of action. The EC/AC-based solutions for the most part do not resort to nonmonotonic reasoning at all, yet fare much better in providing the intuitively warranted inferences than the best-known nonmonotonic approaches. While there are good reasons for ultimately employing nonmonotonic or probabilistic logics – e.g., pervasive uncertainty and the qualification problem – this does show that the scope of monotonic methods has been underestimated. Subsidiary purposes of the paper are to clarify the intuitive status of EC axioms in relation to action effect axioms; and to show how EC, previously formulated within the situation calculus, can be applied within the framework of a temporal logic similar to Sandewall's "discrete fluent logic", with some gains in clarity.

*This is a slight revision, taking account of some criticisms, of a manuscript completed in January 1992 and slated for expansion into a technical report "Explanation closure meets the Sandewall test suite for reasoning about change". The latter will include full proofs, including ones for situation calculus versions of the test problems, but delays in its production have prompted the present abbreviated report.

REPORT DOCUMENTATION PAGE

Form Approved OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*

2. REPORT DATE

3. REPORT TYPE AND DATES COVERED

4. TITLE AND SUBTITLE
Explanation Closure, Action Closure, and the Sandewall Test Suite for Reasoning about Change

5. FUNDING NUMBERS
ONR/DARPA N0014-82-K-0193

6. AUTHOR(S)
Lenhart K. Schubert

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Computer Science Dept.
734 Computer Studies Bldg.
University of Rochester
Rochester NY 14627-0226

8. PERFORMING ORGANIZATION REPORT NUMBER
TR 440

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Office of Naval Research
Information Systems
Arlington VA 22217

DARPA
3701 N Fairfax Dr.
Arlington VA 22203

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Distribution of this document is unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

Explanation closure (EC) axioms were previously introduced as a means of solving the frame problem. This paper provides a thorough demonstration of the power of EC (combined with action closure) for reasoning about dynamic worlds, by way of Sandewall's recently compiled test suite of 12-or-so problems. The problems range from the "Yale turkey shoot" (and variants) to the "stuffy room" problem, and were intended as a test and challenge for nonmonotonic logics of action. The EC/AC-based solutions for the most part do not resort to nonmonotonic reasoning at all, yet fare much better in providing the intuitively warranted inferences than the best-known nonmonotonic approaches. While there are good reasons for ultimately employing nonmonotonic or probabilistic logics--e.g., pervasive uncertainty and the qualification problem--this does show that the scope of monotonic methods has been underestimated. Subsidiary purposes of the paper are to clarify the intuitive status of EC axioms in relation to action effect axioms; and to show how EC, previously formulated within the situation calculus, can be applied within the framework of a temporal logic similar to Sandewall's "discrete fluent logic," with some gains in clarity.

14. SUBJECT TERMS
reasoning about change; frame problem; monotonic reasoning; explanation closure; Sandewayl test suite

15. NUMBER OF PAGES
19

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
unclassified	unclassified	unclassified	UL

1 Introduction

Explanation closure (EC) axioms are complementary to effect axioms. For instance, just as we can introduce effect axioms stating that painting or wallpapering a wall (with appropriate preconditions) changes its color, we can also introduce an EC axiom stating that a change in wall color implies that it was painted or wallpapered. The "closure" terminology signifies that the alternatives given are exhaustive.

This complementarity extends to their use: effect axioms allow the inference of change, and EC axioms the inference of non-change (persistence). For instance, if I know that no-one has painted or wallpapered the wall, then I can conclude that its color has remained unaltered. As first noted by Haas [1987], EC-based persistence reasoning provides a very good handle on the *frame problem*.[1] In [Schubert, 1990] (henceforth Sch90) I extended Haas' work, showing that EC-based techniques generalize to worlds with continuous and agentless change and concurrent actions, and support extremely efficient STRIPS-like methods for tracking effects of successive actions. Moreover, these methods are entirely monotonic as long as the fluents of interest behave relatively simply and deterministically.

In view of their potency, it is surprising that EC-based approaches did not surface much sooner in the history of the frame problem. A commonly expressed qualm about EC axioms is that any enrichment of the (micro)world under consideration is likely to necessitate their revision. For instance, while in a simple world a change in wall color may be attributable to painting or wallpapering, in a more complex world the change may also be due to spraying, tiling, or panelling (or even decay, etc.). True enough - but it is equally true that enrichment of a microworld complicates the effect axioms. For instance, having paint and a brush may be sufficient for successful wall-painting in a simple, benign world, but in a more realistic one, the painter may be thwarted by dried-out paint, an undersize or oversize brush, injury, interference by other agents, etc. (i.e., the *qualification problem* crops up). Yet the fallibility of simple effect axioms has deterred few - not even nonmonotonic theorists - from relying on them! For instance, every formalization of the Yale Turkey Shoot includes axioms asserting that loading a gun makes it loaded, and firing the loaded gun at Fred kills him. This is generally done without comment or apology (except perhaps for a perfunctory gesture toward the qualification problem, which is thereafter ignored). Yet the idea of

[1] A number of other writers have made closely related proposals, e.g., Lansky [1987], Georgeff [1987] Morgenstern & Stein [1988].

turning this around and applying the same strategy to inference of explanations, given a change, seems to occur to almost no-one, and if raised, is met with skepticism.

I am led to believe that there are deep-seated prejudices against the idea of *reasoning deductively against the causal arrow*, perhaps stemming in part from the philosophical tradition on explanation. This tradition holds that physical theories enable us to deduce *resultant* states and events from given ones; while going from results to their causes is not a matter of deduction, but a matter of generating assumptions *from* which we can deduce the results. But while reasoning against the arrow of time and causation (retrodiction, explanation) is apt to generate more alternatives than reasoning with it (prediction), there is no *a priori* physical or logical reason for confining deduction to the forward direction.[2] In view of the the growing evidence for the efficacy of EC reasoning, I would expect these prejudices to be gradually dispelled.

Further reasons for optimism are provided by Reiter's recently developed variant of the EC-based approach to the frame problem [Reiter, 1991]. Rather like Morgenstern & Stein [1988], he focuses on cases where the known effect axioms characterize *all* the ways the changes of interest can come about (Generalized Completeness Assumption). For such cases, he shows how EC axioms can be derived from effect axioms, and combined with them into biconditionals; e.g., a wall changes color *if and only if* it is painted or wallpapered. This mechanical derivation should allay some of the above qualms about the lack of invariance of EC axioms when new actions are added. Reiter further shows how to use such axioms for sound and complete goal regression.

However, I will keep effect axioms and EC axioms separate for the sake of generality, since the Generalized Completeness Assumption fails for some simple worlds of interest. For instance, we may know that a robot's $goto(x)$ action brings about $nextto(Robot, x)$. But it would be wrong to biconditionalize this to say that $nextto(Robot, x)$ becomes true *if and only if* the robot moves next to x. After all, there may be objects near x which the robot may also end up next to (and these "side effects" may depend more or less

[2] It is interesting that people versed in formal logic are apt to regard Sherlock Holmes' "deductions" as misnamed. Rather, they say, Holmes was reasoning inductively or abductively when he constructed explanations for his observations. In my view, if we are willing to grant that the inference of a man's death is deductive, given his unimpeded fall to the pavement from the top of a skyscraper, then some of Holmes' inferences are equally deductive. If the former is not deductive, then no inferences based on world knowledge are deductive, whether directed forward or backward in time.

unpredictably on low-level path planning). Yet we can state an EC axiom that $nextto(Robot, x)$ becomes true *only if* the robot goes to some y and $x = y$ or x is near y; this may be quite sufficient for the persistence reasoning needed for practical purposes (see further details in Sch90). Sandewall's test suite provides additional illustrations [Sandewall, 1991] (henceforth San91). For instance, in the "stuffy room" problem, various EC axioms are possible (without change to the effect axioms), depending on how much freedom to "flit about" we want to allow objects when a vent is blocked or unblocked (creating drafts, one imagines).

The test suite provides an unprecedented opportunity to examine the strengths and shortcomings of various methods for reasoning about change in a systematic way. I will show that the approach based on EC-reasoning fares very well indeed. Moreover, the proffered solutions are monotonic except in the case of one variant of McCarthy's "potato in the tailpipe" problem (where I suggest a probabilistic approach). This seems to me to call for a reassessment of the proper roles of monotonic and nonmonotonic (or probablilistic) methods in reasoning about change. While nonmonotonic methods still retain important role in reasoning about an uncertain, incompletely known world (as the "potato in the tailpipe" problem and other instances of the qualification problem show), monotonic methods can deal straightforwardly with many of the scenarios viewed as motivating examples for nonmonotonic methods.

The examples will also serve to illustrate a version of EC-based reasoning within a temporal calculus loosely modelled on Sandewall's DFL (dynamic fluent logic). They will further illustrate the form and importance of *action closure* (AC) axioms in the temporal calculus, and allow us to probe the limits of the monotonic approach.

2 DFL, TC, and the test scenarios

Sandewall's *discrete fluent logic* (DFL), outlined in a preliminary way in San91, offers a concise notation for time-dependent descriptions of dynamic worlds. A theory of entailment is under development which promises to overcome several shortcomings of extant nonmonotonic logics. The essential semantic idea is this: actions "occlude" the fluents they may affect, for the duration of the action; i.e., the values of occluded fluents cannot be presumed to persist. The model preference criterion is still being refined, but the idea is that less occluded models and those that postpone *transparent* (nonoccluded) change are preferred.

Since the DFL semantics is not yet stable (at the time of this writing), it would be premature to attempt comparison between the behavior of DFL and other logics on the test suite. However, DFL does provide a simple notation which is worth emulating to facilitate future comparisons. More importantly, San91 identifies and catalogues many of the defects of extant nonmonotonic logics, and provides old and new test problems which bring these defects to light. Sandewall's preliminary assessment is that his study "... provides reasons for renewed disappointment. The situation in 1991 is only marginally different from the one in 1986 [the year of the Hanks & McDermott paper]... most of the 'most popular' approaches actually fail on the test scenarios." (*ibid.*: sec. 7).

The "temporal calculus" (TC) notation I will use consists of the usual first-order syntax plus the following temporal notation, largely mimicking DFL (but without involvement of occlusion): Truth of a formula φ at (moment of) time τ is written $[\tau]\varphi$, and truth at all times in $[\tau_1, \tau_2]$ is written $[\tau_1, \tau_2]\varphi$. Another DFL-like notation will be $[\tau_1, \tau_2]\varphi := v$, meaning that $[\tau_1]\neg(\varphi = v)$ and $[\tau_2]\varphi = v$, i.e., the value of φ becomes v somewhere in the interval $[\tau_1, \tau_2]$. If φ is a formula, we use $\varphi = T$ and $\varphi = F$ equivalently with φ and $\neg\varphi$ respectively (as in DFL). As a semantic basis for the notation so far, an interpretation of the fluent predicates and functions is assumed to provide their extensions at each moment of time. (The time line could be taken to be discrete or the real line.) We will also use an *action* predicate *do*, where $[\tau_1, \tau_2]do(\alpha, \beta)$ is true or false of an agent α, action β and time *interval* $[\tau_1, \tau_2]$, viz., the interval over which the action takes place. An interpretation of TC is assumed to specify the extension of *do* at all time *intervals*, rather than at all times. A useful abbreviation will be

$[\tau_1..\tau_2]do(\alpha, \beta)$, which stands for

$(\exists\tau_1')(\exists\tau_2')[\tau_1 \leq \tau_1' \leq \tau_2' \leq \tau_2] \wedge [\tau_1', \tau_2']do(\alpha, \beta)$,

i.e., $do(\alpha, \beta)$ happens *somewhere* between τ_1 and τ_2.

For reasons analyzed in [Schubert, 1993] (henceforth Sch93), the TC solutions to the test problems are generally more perspicuous and concise than solutions in the situation calculus (SC). However, the most interesting difference lies in the way the action closure (AC) assumption – that all relevant actions are known – is encoded. In SC versions, the assumption is implicit in the functional dependence of situations on actions. In TC versions, times (and hence fluent values at those times) are introduced independently of actions, and so the assumption of complete knowledge of relevant actions needs to be stated separately. It will typically (though not always) be represented by the "only if" part of an equivalence of form, "x did y from time t_1 to time t_2 iff (x, y, t_1, t_2) is one of the following tuples...". Such axioms will be called "action chronicles" (with apologies to those, including Sandewall, who have employed the term differently).

An important question here is whether AC assumptions are by their nature excessively strong. Does it not require God-like omniscience to know what all the actions are that could have affected the fluents of interest? The answer is no, provided that we are only looking for *practical* certainty rather than *absolute* certainty. The justification of that answer depends on the origins and purpose of a given scenario. If we are simply being told a *story*, we can rely on the narrator to withhold nothing of relevance from us. The narrator will not neglect to mention that Joe unloaded the gun before pulling the trigger on Fred. As Amsterdam [1991] argues, narrators are expected to tell their story in a way that puts the hearer/ reader on the scene (vicariously, through the narrator's perceptions), and this entails reporting everything of relevance that happened. To be sure, there are many qualifications to be made and subtleties to be explored here. But my point is that the source of closure in narration is the narrator, not the hearer or God. (Formally, Amsterdam assumes that no actions occurred other than those deducible from the narrative, or that could have transpired during explicitly reported lapses in the narrator's awareness. I will have further comments on Amsterdam's proposals later.)

If instead a scenario represents a plan of action, whose consequences are yet to be observed (once the plan is carried out), then clearly it is the planner's *intention* to shield the fluents of interest from capricious disturbances. If you *plan* to kill Fred by loading the gun, aiming at Fred, and pulling the trigger, you surely plan *not* to unload the gun before pulling the trigger. And if you plan to repaint the walls a certain color, you surely do not intend to let others meddle at will. Thus it is the planner who is the source of action closure. He may ensure closure, for instance, by arranging to be the only agent on the scene, or to have only co-agents who will do his bidding, or who at least can be relied on not to interfere. That is all that is needed to justify AC axioms.

Of course, if we demand *absolute* reliability of our axioms, then God-like omniscience is indeed required: after all, even the most carefully insulated and controlled setting is subject to freak occurrences. But that is not an observation about EC or AC axioms in particular, but about *all* nonlogical axioms. For instance, it is often assumed in axiomatizations of Yale-shooting worlds or blocks worlds that loading a gun causes it to be loaded, putting block x on block y causes x to be on y, provided x and y were clear, etc. These *effect* axioms are just as much approximations to reality as EC or AC axioms, in view of the ever-present qualification problem.

Furthermore, I would point out that switching to nonmonotonic methods provides no escape from the "omniscience" criticism. On the contrary, when we assume that some predicate (such as `causes`, `abnormal`, or `clip`) has no elements in its extension other than those dictated by our knowledge, we are making a strong completeness assumption about our knowledge. Furthermore it is hard to tailor these tacit completeness assumptions to what we actually know we know (and don't know), since they involve the undecidable notion of deductive closure. How, for instance, do we express through some sort of predicate minimization that there was no other agent capable of painting, besides the robot, at the scene of the action (for otherwise, I would have seen him), and the robot painted only such-and-such objects at such-and-such times (for otherwise, I would have noticed)? Yet such assumptions are perfectly natural, are a matter of practical certainty – and are easily expressed as EC and AC axioms.

In short, given that all solutions ever offered for Sandewall-like problems have relied on approximation, it is pertinent to ask how far we can get with classical, monotonic approximations alone. Moreover, a monotonic approach to the inference of change and persistence does not preclude the addition of *belief revision* mechanisms, capable of retracting, amending, or adding to the beliefs which form the basis for these monotonic inferences. When I discover that the wall I painted blue turned green when it dried, I'll revise my effect axioms; and if I find that while my back was turned, a prankster who had been hiding in the closet repainted the wall red, I'll revise my action chronicle. But unless and until that happens, I may well be best off reasoning monotonically with "practically certain" axioms.

The test scenarios which follow adhere closely to Sandewall's formulations. Each scenario is described very briefly, the intended conclusions are indicated, and then the TC formalization is shown. Although detailed proofs exist in all cases, confirming that the intended conclusions are reached (Sch93), space limitations prohibit their inclusion. I hope that the axiomatizations are sufficiently transparent to allow the reader to reconstruct the proofs; in one of the more complex cases, the Hiding Turkey Scenario (HTS), a slightly abbreviated proof is included, and this should help with many of the other examples. The headers are worth paying close attention to; they encapsulate essential dimensions of variation among test cases, largely as identified by Sandewall - dimensions often difficult for any one nonmonotonic logic to measure up to simultaneously.

In all of the axiomatizations, names beginning with **obs**, **chr**, **eff**, **exp**, and **ineq** respectively are used for axioms describing observations at particular situations or times, action chronicles, effect axioms, explanation closure axioms, and inequality axioms. These names serve no theoretical purpose, only a mnemonic one (unlike DFL conventions). As in Sch90, constants and functions will start with an upper case letter and variables and predicates will be lower case. Top-level free variables are implicitly universally quantified (with maximal quantifier scope). The predicate u ("unequal") takes any number of arguments and asserts that they are pairwise distinct.

Prediction: Yale Turkey Shoot (YTS)

There are two truth-valued fluents, a (alive) and l (loaded). Initially the turkey is alive and the gun not loaded. The agent loads, waits and fires. Loading brings about l (from prior state $\neg l$ or l), and firing brings about $\neg a$ and $\neg l$ provided that l held prior to it. We wish to conclude that at the end of firing, $\neg a$ holds (the turkey is not alive).

I will slightly embellish the usual action repertoire to include `Unload`, `Spin`, and `Chopneck`, for illustration and for consistency with later variants. For simplicity `Chopneck` (which plays no role here) has been given no preconditions.

obs1 $[0]a \wedge \neg l$

chr1 $[t_1, t_2]do(Joe, y) \Leftrightarrow (t_1, t_2, y) \in \{(4, 6, Load), (10, 12, Fire)\}$

eff1 $[t_1, t_2]do(Joe, Load) \Rightarrow [t_2]l$

eff2 $[t_1]l \wedge [t_1, t_2]do(Joe, Fire) \Rightarrow [t_2](\neg a \wedge \neg l)$

eff3 $[t_1, t_2]do(Joe, Chopneck) \Rightarrow [t_2]\neg a$

exp1 $[t_1, t_2]l := T \Rightarrow [t_1..t_2]do(Joe, Load)$

exp2 $[t_1, t_2]l := F \Rightarrow (\exists y \in \{Fire, Unload, Spin\})[t_1..t_2]do(Joe, y)$

exp3 $[t_1, t_2]a := F \Rightarrow (\exists t_1')[t_1 \leq t_1' \leq t_2 \wedge [t_1']l \wedge [t_1'..t_2]do(Joe, Fire)]$
$\vee [t_1..t_2]do(Joe, Chopneck)$

ineq1 $u(Load, Unload, Fire, Spin, Chopneck)$

Though superficially close to Sandewall's axiomatization, the TC version makes significantly stronger assumptions at the outset. For instance, `chr1` leaves Joe inactive between loading and firing, and this together with `exp2` ensures that the gun remains loaded. But in the DFL version, this is a defeasible chronicle completion inference. Should it be? Suppose the problem specification included the statement, "Between loading and firing, another action either did or did not take place". Intuitively, this blocks the inference that the gun remained loaded – despite the fact that the added statement is logically vacuous (a *tautology*)!

Clearly, it is a mistake to simply render the given English sentences as directly as possible in some logic, and then make it a matter of the semantics of that logic to deliver the intuitively required conclusions. How could *any* reasonable logic have entailments defeasible by tautologies? This once again raises the important question of "what's in a problem statement". As noted earlier, Amsterdam [1991] drew attention to the role of narrative conventions in story-like problem statements, in particular the requirement that the author relate everything his audience would have observed under the reported circumstances – except perhaps events that transpired during explicitly reported lapses of attention (e.g., where the author indicates that some time passed, or says "I blacked out for a moment", etc.) This is formally written as $\mathtt{U}A_t$, i.e., it is unknown whether action A occurred.

It is interesting to note that Amsterdam's assumption about what actions did and did not occur is closely related to the AC assumption. Stated a little more fully than before, his assumption is that an action A occurred at t if A_t is provable, and did not occur if neither A_t nor UA_t is provable. My AC assumption is computationally less problematic: it says that all the actions that bear on the fluents of interest are explicitly known, without invoking provability. Also, Amsterdam makes an assumption closely related to EC: roughly speaking, changes that are provable effects of provable actions (according to some theory of what constitutes an "effect") definitely occurred, and no change occurred unless it is the effect of some A_t, where A_t or UA_t is provable. (For the exact formulation, see [1991].) Amsterdam notes that his approach fails to allow for actions which people regard as "obvious" inferences from certain state changes. (His example is one where a character is sitting by the fireplace in one sentence and standing by the door in the next.) These are precisely the action inferences supplied by EC!

Amsterdam's attempt to capture narrative conventions by nonmonotonic action and effect closure and the modal U operator is interesting, but it remains to be seen how far it can be taken. Besides computational intractability and the problem about action inference noted by Amsterdam, there is also the problem that real stories allow for many actions and events that are neither entailed by the story nor occluded by lapses in the narrator's attention. For instance, it certainly seems possible in a story like *Little Red Riding Hood* that the heroine hopped over a small creek, or glanced at some birds overhead on her way to Grandmother's house, even though nothing in the story entails this or even suggests that this *may* have occurred. The narrator simply did not judge such events relevant, and therefore, abiding by the Gricean maxim, omitted them. The view taken here is that narrative implicatures and domain reasoning are separable phenomena, and that it is therefore worthwhile to study domain reasoning methods as far as possible independently of story understanding. This means that we begin by extracting *all* of the information intuitively conveyed by a narrative – the positive as well as the negative, the asserted as well as the "conversationally implicated" information – while setting aside the question of exactly *how* the narration managed to convey that information. Only then do we ask what *follows* from what we have been told.

Regardless of strategy, however, what is important about Amsterdam's work is its recognition of the importance of narrative conventions and maxims in shaping what we take a story to imply. Much of the heated debate about which nonmonotonic logic is the right one for chronicle completion seems attributable to the neglect of information implicitly conveyed through these conventions and maxims, or misguided attempts to make this information fall out of the logic.

Retrodiction: the Stanford Murder Mystery (SMM)

The world is the same as for the YTS, but the gun is initially loaded, firing and waiting are performed in succession, and then the turkey is not alive. We are to infer that the gun was initially loaded, and the turkey was not alive after firing (prior to the wait).

obs1 $[0]a$

chr1 $[t_1, t_2]do(Joe, y) \Leftrightarrow (t_1, t_2, y) = (10, 12, Fire)$

obs2 $[14]\neg a$

eff1-ineq1 as above (YTS)

Ambiguous prediction: the Ferryboat Connection Problem (FCP)

A motorcycle M goes from F, some location on island Fyen, to the ferry landing L, and gets there between times 99 and 101. If it gets there before time 100, it will catch the ferry and be in Jutland (J) as of time 110, otherwise it stays at L. We are to infer that at time 110, M is either on L or on J (but should not infer one or the other).

Actually, Sandewall's DFL formalization makes the problem a little harder by saying, in effect:

At time 0, the bike is on Fyen. At some time T between 99 and

101, the bike arrives at the landing. If its arrival T is before time 100, then the bike gets on board the ferry at time 100. If the bike is on board at time 105, it arrives on Jutland at time 110.

I will use a similar encoding for the TC version.

obs1 $[0]on(M, F)$

chr1 $[t_1, t_2]do(M, y) \Leftrightarrow (t_1, t_2, y) = (0, T, GotoL)$
$\vee[T < 100 \wedge (t_1, t_2, y) = (100, 101, Board)]$
$\vee[T \geq 100 \wedge T \leq t_1 \leq t_2 \leq 110 \wedge y = Wait]$

obs2 $99 \leq T \leq 101$

eff1 $[t_1, t_2]do(M, GotoL) \Rightarrow [t_2]on(M, L)$

eff2 $[t_1, t_2]do(M, Board) \Rightarrow [t_2]on(M, B)$

eff3 $[105]on(M, B) \Rightarrow [110]on(M, J)$

eff4 $[t_1, t_2]do(M, Wait) \wedge [t]on(M, y) \Rightarrow [t_1, t_2]on(M, y)$

expl $[t_1, t_2]on(M, B) := F \Rightarrow [t_1..t_2]do(M, Unboard)$

ineq1 $u(GotoL, Board, Unboard, Wait)$

This assumes somewhat more than Sandewall's DFL version, in its use of `Unboard`. The EC axiom giving `Unboard` as an explanation for $on(M, B)$ becoming false (i.e., **expl**) is not just an embellishment but is essential to the inference that once M in on board, it stays on board till time 110.

I claim that this is an entirely reasonable assumption – in fact, that the desired conclusion about ending up at L or J should *not* be reached based on the information assumed by Sandewall. To resolve this point, I haphazardly contrived several "stories" satisfying Sandewall's axioms (about a man going to dinner, a subway trip, a house fire, etc.) and found that for more than half of them the disjunctive conclusion "at L or J" was unwarranted. (Three of the "stories" are in Sch93.) So once again, linguistic pragmatics appears to have intruded into the analysis.

Prediction from disjunction: the Russian Turkey Shoot (RTS)

The problem differs from YTS only in that a `Spin` action (spinning the chamber of the gun) is inserted between the `Wait` and the `Fire`. The inference that the turkey dies should be disabled.

chr1 $[t_1, t_2]do(Joe, y) \Leftrightarrow$
$(t_1, t_2, y) \in \{(4, 6, Load), (7, 9, Spin), (10, 12, Fire)\}$

obs1, eff1-3, expl-3, ineq1 as in YTS

Ambiguous retrodiction: Stolen Car Problem (SCP)

At the beginning of the first night, the car is in my possession (expressed by predicate p). I perform the action of "leaving the car overnight in my garage" on two successive nights. On the following evening, the car is not in my possession.

I cannot lose possession of the car during the day. Once I've lost possession of it, I can't regain it. The intended conlusion is that I lost possession of the car during one of the two nights (with no conclusion about which night it was).

To illustrate that complete action closure is in general unnecessary, I will merely assume that the only `Leave-car-overnight` actions were those on the given nights ($[0,2]$ and $[4,6]$), so (given that only these can lead to car loss) the car couldn't have been lost during the day. The even weaker assumption that there were no `Leave-car-overnight` actions on the given days would have been sufficient, as well.

obs1 $[0]p$

chr1 $[t_1, t_2]do(I, Leave\text{-}car\text{-}overnight) \Rightarrow (t_1, t_2) \in \{(0,2), (4,6)\}$

obs2 $[8]\neg p$

eff1 $[t_1]\neg p \Rightarrow [t_1, t_2]\neg p$

exp1 $[t_1, t_2]p := F \Rightarrow [t_1..t_2]do(I, Leave\text{-}car\text{-}overnight)$

Logically related fluents: Dead Xor Alive Problem (DXA)

This is a slight reformulation of the YTS, with "becoming not alive" replaced by "becoming dead", and the equivalence axiom $[t]\neg a \Leftrightarrow [t]d$ added (where d means "dead"). Such logical connections lead to "autoramifications" (in Sandewall's terminology). In our monotonic approach the reformulation leads unproblematically to the conclusion that the turkey is d (and hence $\neg a$) after firing, and no sooner, much as before.

Logically related fluents: Walking Turkey Problem (WTP)

This is another slight variant of the YTS, in which the turkey is initially known to be walking (w) (but it is not explicitly given that he is alive), and the conditional $[t]w \Rightarrow [t]a$ is known. We are to conclude that the turkey is not walking after the firing. We easily infer $[0]a$ from $[0]w$ and reason as in YTS, concluding $[10]a$ and $[12]\neg a$ and hence $[12]\neg w$ by the contrapositive of the new conditional.

Prediction from disjunction: Hiding Turkey Scenario (HTS)

In this variant of Sandewall's, the turkey may or may not be deaf, and if it is not, it goes into hiding when the gun is loaded (where it is initially unhidden). Gun-loading, waiting, and firing take place in succession as in the YTS, but firing only kills the turkey if it is not hiding.

The intended conclusion is that at the end of firing, the turkey is either deaf and not alive, or nondeaf and alive. Sandewall points out that this problem confutes methods like Kautz's [1986] which unconditionally prefer later changes to earlier ones (and so leave the turkey unhidden and hence deaf and doomed). In an EC-based approach, this variant is quite analogous to the RTS. We add an effect axiom that `Hide` brings about h (**eff4**), and EC axioms that *only* `Hide` and `Unhide` can bring about h and $\neg h$ respectively (**exp3**, **exp4**). We further add a "once deaf, always deaf" assumption (**eff5**) and take "ceasing to be alive" as the only explanation for becoming deaf (**exp5**).

I will represent the gunman's (Joe's) and the turkey's (Fred's) actions by separate chronicles for clarity (`chr1` and `chr2`).

obs1 $[0]a \wedge \neg l \wedge \neg h$

chr1 $[t_1, t_2]do(Joe, y) \Leftrightarrow (t_1, t_2, y) \in \{(4, 6, Load), (10, 12, Fire)\}$

chr2 $[t_1, t_2]do(Fred, y) \Leftrightarrow [[5]\neg d \wedge (t_1, t_2, y) = (7, 9, Hide)]$

eff1 $[t_1, t_2]do(Joe, Load) \Rightarrow [t_2]l$

eff2 $[t_1](l \wedge \neg h) \wedge [t_1, t_2]do(Joe, Fire) \Rightarrow [t_2](\neg a \wedge \neg l)$

eff3 $[t_1, t_2]do(Joe, Chopneck) \Rightarrow [t_2]\neg a$

eff4 $[t_1, t_2]do(Fred, Hide) \Rightarrow [t_2]h$

eff5 $[[t_1]d \wedge t_2 \geq t_1] \Rightarrow [t_2]d$

exp1 $[t_1, t_2]l := F \Rightarrow (\exists y \in \{Fire, Unload, Spin\})[t_1..t_2]do(Joe, y)$

exp2 $[t_1, t_2]a := F \Rightarrow (\exists t_1')[t_1 \leq t_1' \leq t_2 \wedge [t_1'](l \wedge \neg h)$
$\wedge[t_1'..t_2]do(Joe, Fire)] \vee [t_1..t_2]do(Joe, Chopneck)$

exp3 $[t_1, t_2]h := T \Rightarrow [t_1..t_2]do(Fred, Hide)$

exp4 $[t_1, t_2]h := F \Rightarrow [t_1..t_2]do(Fred, Unhide)$

exp5 $[t_1, t_2]d := T \Rightarrow [t_1..t_2]a := F$

ineq1 $u(Load, Unload, Fire, Spin, Chopneck, Hide, Unhide)$

Reasoning: Suppose the turkey is initially deaf, $[0]d$. Then he is still deaf after the `Load` by `eff5`, hence he fails to `Hide` after the `Load` (or indeed, at any time) by `chr2`. Since he is initially unhidden according to `obs1`, he remains unhidden by `exp3`(etc.), so that in particular $[10]\neg h$. Likewise the l property inferrable at time 6 from the `Load` action and `eff1` persists by EC-reasoning to time 10. Hence the `Fire` action is fatal by `eff2`, and so $[12]\neg a$ and $[12]d$ (after another application of `eff5`).

On the other hand, if the turkey is not initially deaf, he is still nondeaf at time 5 during the `Load` (for otherwise he would have become nonalive between times 0 and 5 by **exp5**, and hence there would have had to be a `Fire` or `Chopneck` action by Joe between those times, contrary to `chr1`.) Hence Fred `Hides` during $[7,9]$ by `chr2`. He remains hidden through the subsequent actions by EC-reasoning based on **exp4**, in particular $[10]h$. After proving persistence of a from the initial state to time 10 in the usual way, we can also prove its persistence through the `Fire` action from **exp2**. Thus $[12]a$ and $[12]\neg d$ in this case (after further applications of **exp5** and **exp2**). The assumption of initial deafness or non-deafness can each be made consistently, so that we can only infer the disjunction of the corresponding conclusions.□

Improbable disturbances: Potato in the Tailpipe (TPP)

Initially the car engine is not running ($\neg r$). The action of attempting to start the car is performed. On the assumption that there is usually no potato is the tailpipe (predicate p is usually false), and that the car will start if there isn't, we are to conclude that the car will start.

Sandewall approximates the premise that there is usually no potato in the tailpipe by saying that there is no potato in the tailpipe at time 0, by default. Default axioms are used only for final ranking of the most preferred models of the remaining axioms, and thus may be violated.

Ordinary first-order logic cannot express explicitly uncertain premises (such as ones involving "usually") and so cannot accurately model reasoning based upon them. To my mind the most attractive approach to uncertain reasoning is one based on *direct inference* of epistemic probabilities (i.e., probabilities for particular propositions) from "statistical" generalizations (e.g., [Kyburg, 1988], [Bacchus, 1988; Bacchus, 1990], [Halpern, 1990]). The advantage of a probabilistic approach to nonmonotonicity is that it allows systematically for degrees of belief, and that it can provide a coherent basis for decision-making by an intelligent agent. The present version of the TPP provides only a trivial illustration of the probabilistic approach, but more "realistic" variants can be found in [Tenenberg, 1991], [Tenenberg and Weber, 1992] and [Bacchus, 1990][180-2].)

The following, then, is a TC-like axiomatization of TPP based on a statistical interpretation of the statement about potatoes in the tailpipe.

stat1 $[[t]\neg p]_t > .99$

obs1 $[0]\neg r$

chr1 $[t_1, t_2]do(Joe, y) \Leftrightarrow (t_1, t_2, y) = (6, 8, Start)$

eff1 $[t_1]\neg p \wedge [t_1,t_2]do(Joe, Start) \Rightarrow [t_2]r$

Reasoning: From `stat1` & `eff1`, $[[t_2]r|[t_1,t_2]do(Joe, Start)]_{t_1} > .99$ (t_2 is still universally quantified). Hence by `chr1`, $Prob([8]r) > .99$, using universal instantiation for t_2 and direct inference for $t_1 = 6$.□

As long as we demand that very improbable, but nevertheless possible. events be explicitly allowed for, the TPP cannot be monotonically represented. Still, the following trivial approximation is worth noting. Here the tailpipe is assumed to be initially clear, and the assumed chronicle and EC axiom for tailpipe plugging-up rule out any mischief.

obs1 $[0]\neg r$

obs2 $[0]\neg p$

chr1 $[t_1,t_2]do(x,y) \Leftrightarrow (t_1,t_2,x,y) = (6,8,Joe,Start)$

eff1 $[t_1]\neg p \wedge [t_1,t_2]do(Joe, Start) \Rightarrow [t_2]r$

exp1 $[t_1,t_2]p := T \Rightarrow (\exists x)[t_1..t_2]do(x, Plug)$

Event-time ambiguity: the Tailpipe Marauder (TPM)

This variant of Sandewall's assumes that a potato *is* put in the tailpipe somewhere between 8am and 5pm, but it is not known when. The attempt to start the car takes place at 1:30pm, and the aim is *not* to reach a conclusion about whether the car starts or not.

TPM is much less problematic for a monotonic approach than the original TPP, since it merely involves *incomplete* knowledge (about the time of a known event), rather than "defeasible" knowledge (where one of the possibilities consistent with our incomplete knowledge is much more probable than the others). I'll arbitrarily call the protagonist Joe and the antagonist Moe, and assign a duration of 2 (two hundredths of an hour) to the `Plug` and `Start` actions.

obs1 $[0]\neg r$

obs2 $[0]\neg p$

chr1 $[t_1,t_2]do(x,y) \Leftrightarrow (t_1,t_2,x,y) \in$
$\{(1350,1352,Joe,Start),(T,T+2,Moe,Plug)\}$,
$800 \leq T \leq 1699$

eff1 $[t_1,t_2]do(x,Plug) \Rightarrow [t_2]p$

eff2 $[t_1,t_2]do(x,Start) \Rightarrow [[[t_1]p \Rightarrow [t_2]\neg r] \wedge [[t_1]\neg p \Rightarrow [t_2]r]]$

exp1 $[t_1,t_2]p := T \Rightarrow [t_1..t_2]do(Moe, Plug)$

exp2 $[t_1,t_2]p := F \Rightarrow [t_1..t_2]do(Joe, Unplug)$

exp3 $[t_1,t_2]r := T \Rightarrow [t_1..t_2]do(Joe, Start)$

ineq1 $u(Start, Plug, Unplug)$

It is straightforward to show that neither $[1352]\neg r$ nor $[1352]r$ can be inferred. Note that if we are *given* $[1352]\neg r$, we can infer $T < 1350$ and if we are *given* $[1352]r$, we can infer $T \geq 1350$.

Event-order ambiguity: Tailpipe Repairman Scenario (TPR)

In this variant, Sandewall assumes that the tailpipe is initially blocked, and the actions of unplugging the tailpipe and trying to start the car are done in arbitrary order. No action ordering should be inferrable, but it should follow that the car starts iff the unplugging is done first.

I include the gratuitous assumption that the tailpipe was unobstructed prior to 8am, for conformity with Sandewall's axiomatization.

obs1 $[800]\neg r$ (not running at 8am)

obs2 $[0]\neg p$ (no potato previous midnight)

obs3 $[800]p$ (potato in tailpipe at 8am)

chr1 $[[t_1, t_2]do(x, y) \wedge (800 \leq t_1 \leq 1698) \wedge (800 \leq t_2 \leq 1698)]$
$\Leftrightarrow (t_1, t_2, x, y) \in \{(T_1, T_1{+}2, Joe, Start), (T_2, T_2{+}2, Joe, Unplug)\}$

eff1 $[[t_1]\neg p \wedge [t_1, t_2]do(Joe, Start)] \Rightarrow [t_2]r$

eff2 $[t_1, t_2]do(Joe, Unplug) \Rightarrow [t_2]\neg p$

exp1 $[t_1, t_2]p := F \Rightarrow (\exists x)[t_1..t_2]do(x, Unplug)$

exp2 $[t_1, t_2]p := T \Rightarrow (\exists x)[t_1..t_2]do(x, Plug)$

exp3 $[t_1, t_2]r := T \Rightarrow (\exists t_1' \geq t_1)(\exists x)[t_1']\neg p \wedge [t_1'..t_2]do(x, Start)$

ineq1 $u(Start, Plug, Unplug)$

Neither $[T_1{+}2]r$ nor $[T_1{+}2]\neg r$ can be inferred. With assumption $T_2{+}2 \leq T_1$, we would get $[T_1 + 2]r$, and for the contrary assumption we find the car will never run. Given the extra premise `obs2`, it is also possible to deduce a `Plug` action prior to 8am.

Stable and unstable worlds: additional scenarios

San91 mentions plans for including some more complex scenarios in the test suite, in particular Lifschitz's N-blocks world [Lifschitz, 1987], and Winslett's variants of Ginsberg & Smith's "stuffy room" scenario [Winslett, 1988]. It will be interesting to see how various "rival" nonmonotonic logics (other than those of the responsible authors) fare in these slightly more intricate worlds. As far as the EC-based approach is concerned, they present no unusual challenge (and indeed at least equally complicated cases were treated in Sch90).

The N-blocks world allows movement of one block onto another (with the usual clear-top conditions, formulated in a slightly unusual way in terms of a *top* function) or onto the table, and painting of a block with one of three colors. There are axioms about uniqueness of destinations and resultant colors, and so on. The aim is to come up with the same state-transition characterization of this world as Lifschitz obtains circumscriptively, i.e., (roughly) nothing else changes when a block is moved or painted. In an EC-approach, one adds straightforward EC axioms about the 5 fluent predicates employed $(at, color, true, false, clear)$: blocks change *at* properties only when moved, and change color only when painted, etc. In fact, Reiter's technique for automatic biconditionalization should work well here.

Lifschitz's world is as stable as one would expect a blocks world to be, whereas the "stuffy room" world leaves considerable room for objects "flitting about" capriciously when another is moved. The intuitive picture which is intended to support this degree of instability is based on floor ducts, drafts, and light objects like newspapers. While Winslett's "possible models approach" does seem to fare better than the more syntactic "possible worlds approach" it was intended to correct, it is curiously selective in the motions it sanctions. For instance in the first problem the birdcage may flit from one duct to the other, but not to the floor; moreover, the floor ducts seem to act as "attracters" *because* their being blocked keeps the room stuffy! In an EC approach, one can model a Lifschitz-like stable world, or one in which, say, any lightweight object can shift when things are placed on ducts or moved away, or even ones in which ducts act as "attracters", or for that matter, "repellers". Thus we are able to tailor the persistence knowledge to fit the physics, and this strikes me as more sensible than trying to make the physics fall out of the semantics.

3 Conclusion

San91 ends on a rather pessimistic note, and speculates that part of the problem may be that many nonmonotonic logics have evolved largely in conference proceedings, whose format precludes thorough evaluation and comparison. I have tried in this report to explore the scope of a particular technique, EC/AC-based reasoning in dynamic worlds, more fully than is the standard practice. I hope to have provided enough of the technical gist of the proposed EC/AC-based solutions to Sandewall's test suite to support my contention that much of the reasoning commonly thought to require nonmonotonic methods can in fact be done monotonically.

I should reiterate that in saying this, I am not suggesting that monotonic reasoning is all you really need. A monotonic theory of any realistically complex, dynamic world is bound to be an approximation, in the sense that it ignores both improbable qualifications on the effects of actions, and far-fetched explanations for change. We simply cannot *express* in ordinary FOL that certain kinds of events are very unlikely but may nonetheless occur. For this, we need to go beyond FOL, as has been done in nonmonotonic and probabilistic logics.

But I think the literature on nonmonotonic logics has put *too much* of the burden of commonsense reasoning (especially too much of the task of inferring persistence and change) on nonmonotonic methods. An adverse effect has been a confusion between narrative principles and logic, and between physics and logic. The very terms "persistence" and "inertia" used as *model-theoretic* notions may be inapt, since objects stay put, or keep moving, for physical rather than model-theoretic reasons. As well, the over-deployment of nonmonotonic methods has created computational intractability problems, where relatively simple monotonic methods would have sufficed.

The EC/AC-based approach seems to deal with most of the issues addressed by Sandewall's test suite rather handily. It does not render things quiescent (or nonexistent) merely because nothing is known about them, it does not spawn spurious events to minimize change, it does not fail when aimed backward in time, and it does not arbitrarily choose between disjuncts. Plausible EC and AC axioms are not hard to conjure up (and as Reiter showed, the former can sometimes be obtained mechanically), they do not work in mysterious ways, and they work computably and even efficiently (in STRIPS-like settings). It therefore seems well worthwhile to further investigate EC/AC-based methods, e.g., for planning applications. One of the most interesting directions for further work is to use probabilistically qualified EC and AC axioms in a probabilistic logic setting (cf. the earlier citations of work by Bacchus and Tenenberg & Weber); i.e., we would say such-and-such a change is *very likely* due to this or that kind of action, and such-and-such actions are *very probably* the only relevant ones that occurred in a certain setting. At that point we would be ready to address the qualification problem in full, while still exploiting the power of EC and AC to infer (probable) persistence or change.

Acknowledgements

Conversations with Ray Reiter and Andy Haas have clarified my understanding of the relation between effect axioms, EC axioms, and the qualification problem. Support was provided by ONR/DARPA research contract no. N00014-82-K-0193 and Rome Lab contract F30602-91-C-0010.

References

[Amsterdam, 1991] Amsterdam, J. 1991. Temporal reasoning and narrative conventions. In *Proc. of the 2nd RInt. Conf. on Principles of Knowledge Representation and Reasoning (KR'91)*, Cambridge, MA. 15-21.

[Bacchus, 1988] Bacchus, F. 1988. Statistically founded degrees of belief. In *Proc. of the 7th Bienn. Conf. of the Can. Soc. for Computational Stud. of Intelligence (CSCSI '88)*, Edmonton, Alberta. 59-66.

[Bacchus, 1990] Bacchus, F. 1990. *Representing and Reasoning with Probabilistic Knowledge*. MIT Press, Cambridge, MA.

[Brown, 1987] Brown, F. M., editor 1987. *The Frame Problem in Artificial Intelligence. Proc. of the 1987 Workshop.* Lawrence, KS. Morgan Kaufmann Publishers, Los Altos, CA.

[Georgeff and Lansky, 1987] Georgeff, M.P. and Lansky, A.L., editors 1987. *Reasoning about Actions and Plans: Proc. of the 1986 Workshop.* Timberline, OR. Morgan Kaufmann Publ., Los Altos, CA.

[Georgeff, 1987] Georgeff, M. P. 1987. Actions, processes, causality. In *M.P. Georgeff and A.L. Lansky (1987)*. 99-122.

[Haas, 1987] Haas, A.R. 1987. The case for domain-specific frame axioms. In *F. M. Brown (1987)*. 343-348.

[Halpern, 1990] Halpern, J.Y. 1990. An analysis of first-order logics of probability. *Artificial Intelligence* 46:311- 350.

[Kautz, 1986] Kautz, H. 1986. The logic of persistence. In *Proc. of the 5th Nat. Conf. on AI (AAAI 86)*, Philadelphia, PA. 401-405.

[Kyburg, 1988] Kyburg, H. 1988. Probabilistic inference and probabilistic reasoning. In Shachter, and Levitt, , editors 1988, *The Fourth Workshop on Uncertainty in Artif. Intell.* 237-244.

[Lansky, 1987] Lansky, A.L. 1987. A representation of parallel activity based on events, structure, and causality. In *M.P. Georgeff and A.L. Lansky (1987).* 123–159.

[Lifschitz, 1987] Lifschitz, V. 1987. Formal theories of action. In *F. M. Brown (1987).* 35–57.

[Morgenstern and Stein, 1988] Morgenstern, L. and Stein, L.A. 1988. Why things go wrong: a formal theory of causal reasoning. In *Proc. of the 7th Nat. Conf. on AI (AAAI 88)*, Saint Paul, MN. 518–523.

[Reiter, 1991] Reiter, R. 1991. The frame problem in the situation calculus: a simple solution (sometimes) and a completeness result for goal regression. In V. Lifschitz (ed). *Artificial Intelligence and Mathematical Theory of Computation*. Academic Press. 359–380.

[Sandewall, 1991] Sandewall, E. 1991. Features and fluents. *Dept. of Computer Science. Res. Rep.* (ISSN-0281-4250).

[Schubert, 1990] Schubert, L. K. 1990. Monotonic solution of the frame problem in the situation calculus: an efficient method for worlds with fully specified actions. In H. Kyburg, R. Loui and G. Carlson (Eds.) *Knowledge Representation and Defeasible Reasoning* 23–67.

[Schubert, 1993] Schubert, L. K. 1993. Explanation closure and action closure meet the sandewall test suite. Dept. of Computer Science, Univ. of Rochester, Technical Report (to appear).

[Tenenberg and Weber, 1992] Tenenberg, J. and Weber, J. 1992. A statistical approach to the qualification problem. Dept. of Computer Science, Univ. of Rochester. Technical Report 397.

[Tenenberg, 1991] Tenenberg, J. 1991. Abandoning the completeness assumption: a statistical approach to solving the frame problem. *Int. J. of Expert Systems* 3(3,4).

[Winslett, 1988] Winslett, M. 1988. Reasoning about action using a possible models approach. In *Proc. of the 7th Nat. Conf. on AI (AAAI 88)*. Saint Paul, MN. 89–93.